Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms

M. HAMADA, H. F. MARTZ, C. S. REESE, and A. G. WILSON

This article shows how a genetic algorithm can be used to find near-optimal Bayesian experimental designs for regression models. The design criterion considered is the expected Shannon information gain of the posterior distribution obtained from performing a given experiment compared with the prior distribution. Genetic algorithms are described and then applied to experimental design. The methodology is then illustrated with a wide range of examples: linear and nonlinear regression, single and multiple factors, and normal and Bernoulli distributed experimental data.

KEY WORDS: Expected information gain; Logistic regression; Linear and nonlinear regression; Multifactor designs; Shannon information.

1. INTRODUCTION

This article presents and illustrates a practical, easy-to-use technique for obtaining near-optimal Bayesian experimental designs for regression models. The technique is based on the use of a genetic algorithm (GA), and the designs we seek are those that nearly maximize the expected gain in Shannon information provided by the experiment. We illustrate the broad applicability of our approach using five examples, which include both linear and nonlinear models as well as continuous and binary responses.

As in many other areas of statistics, in the past few decades we have seen a significant increase in Bayesian methods in experimental design for regression models. A major reason for this interest is that, before an experiment is conducted, pertinent information is often available that can be formally considered in a Bayesian approach. In fact, the existence of this "prior" information often serves as a prime motivation for the experiment. Chaloner and Verdinelli (1995) gave an excellent overview of Bayesian experimental design for both regression and analysis of variance models.

Following earlier decision analysis work by Raiffa and Schlaifer (1961), Lindley (1972) suggested a decision-theoretic approach to Bayesian experimental design. For a specified utility function that reflects the purpose of the experiment, he suggested that a design be chosen that maximizes the expected utility. Here the expectation is taken with respect to the two classes of unknowns: the sample response data, which has yet to be observed when the design is being considered, and the unknown values of the parameters in an assumed response model. A design which maximizes the expected utility is known as an "optimal" Bayesian experimental design.

The choice of a utility function is extremely important. Lindley (1956) suggested that the expected Shannon information gain (Shannon 1948) might be a useful utility, and Stone (1959), DeGroot (1962, 1986), Bernardo (1979) and others have followed Lindley's suggestion to choose designs that maximize the expected gain in Shannon information provided by the experiment. We note that this criterion is equivalent to choosing designs that maximize the expected Kullback–Leibler distance between the prior and posterior distributions (Chaloner and Verdinelli 1995).

In the well-known case of a normal linear regression model $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\mathbf{R})$, and $\sigma^2, \boldsymbol{\beta}_0$ and $\mathbf{R}$ are known, maximization of the expected Shannon information gain is equivalent to choosing a design that maximizes the determinant of the sum of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{R}^{-1}$. Maximizing this determinant is known as the Bayesian D-optimality criterion. Bayesian D-optimality for linear regression models was considered by Stone (1959), Sinha (1970), Guttman (1971), Smith and Verdinelli (1980), Dette (1993a, 1993b), and Verdinelli (2000). More recently, Dette and Sperlich (1994, 1996), Mukhopadhyay and Haines (1995), He, Studden, and Sun (1996), Dette (1996), Dette and Neugebauer (1997), AndereRendon, Montgomery, and Rollier (1997), Dette and Wong (1998), Song and Wong (1998), and Haines (1998) have considered Bayesian D-optimal designs in nonlinear regression models.

Unfortunately, in many practical cases (such as when the variance σ^2 in the linear regression model considered above is unknown), the integral defining the expected Shannon information gain is intractable. Thus, calculating the expected Shannon information gain (the utility), as well as maximizing it to obtain the desired optimal Bayesian design, are mathematically difficult tasks. This difficulty has had two major consequences.

First, with the exception of Flournoy (1993) and Clyde, Muller, and Parmigiani (1996), few articles have appeared in which Bayesian methods have actually been used to determine optimal experimental designs prior to performing the actual experiment.

Second, various numerical methods for determining "approximately optimal" Bayesian experimental designs have been pro-

M. Hamada, H. F. Martz, C. S. Reese, and A. G. Wilson are Technical Staff Members in Statistical Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545 (E-mail: hamada@lanl.gov). We thank Ken Ryan for his comments on an earlier version and Dee Won for her encouragement of this work. We also thank the associate editor and two anonymous referees for their insightful and helpful comments that greatly improved this manuscript.

posed. These approximations include methods that involve simulation, those that approximate the true posterior distribution (e.g., using a normal distribution), those that approximate the prior (such as with a discrete approximation), and those that approximate the marginal distribution (such as using Laplace's method). Pilz (1991) and Chaloner and Verdinelli (1995) discussed these and other approximate methods for both linear and nonlinear models. Muller (1998) also reviewed several simulation-based methods for estimating the expected gain in Shannon information and the use of simulated annealing for determining optimal Bayesian designs.

GAs are stochastic optimization methods that use Darwinian models of population biology for obtaining near-optimal solutions to multivariable objective functions. They are ideally suited for searching irregular, poorly characterized function spaces. The number of variables simultaneously considered by GAs may range from a few through hundreds (sometimes even thousands). GAs are extremely flexible in that they do not require the usual mathematical restrictions of strict continuity, differentiability, convexity, and so on of the objective function. The variables can be some combination of continuous, discrete, or categorical variables, and the continuous variables may also be ordered. Thus, for reasons such as these, GAs have become quite useful in practice. Goldberg (1989), Michalewicz (1992), and Holland (1992a) are excellent textbooks on GAs, while Holland (1992b) provides a nice introductory tutorial.

GAs have only recently been considered in statistical applications. Broudiscou, Leardi, and Phan-Tan-Luu (1996) used a GA to construct standard D-optimal designs, and provided a nice general introduction to the use of GAs in design. Chatterjee, Laudato and Lynch (1996) introduced the use of GAs in a broad range of statistical applications. In that article they concluded "many statistical and mathematical restrictions that usually restrict modeling and analysis can be dispensed with by employing the GA as an optimization technique." Following Taguchi's robust design ideology, Forouraghi (2000) used a GA to obtain multiobjective robust designs.

We use a GA to determine near-optimal Bayesian experimental designs for a broad class of regression models. The class includes both linear and nonlinear models as well as both continuous and binary responses. For convenience, we restrict consideration to continuous independent (or predictor) variables, although the extension to categorical factors (such as ANOVA models) is straightforward.

Section 2 presents a practical and easy-to-apply GA for use in solving the Shannon expected information gain criterion to obtain near-optimal Bayesian experimental designs for regression models. Section 3 illustrates the method using a broad range of examples. Finally, Section 4 presents a summary and some conclusions.

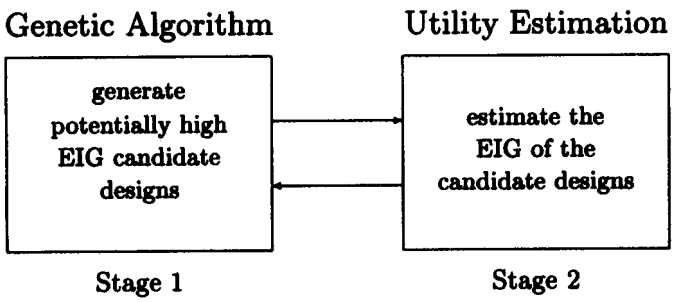

Figure 1. Two-stage Iterative Bayesian Experimental Design Solver

2. NEAR-OPTIMAL BAYESIAN EXPERIMENTAL DESIGNS

For a given experimental design $\mathbf{X}$, data $\mathbf{y}$ is observed according to a specified sampling model $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$. Here $\boldsymbol{\theta}$ is a vector of unknown parameters, of interest to be estimated, in the sampling model. Although the prior distribution $\pi(\boldsymbol{\theta})$ does not depend on $\mathbf{X}$, the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$ depends on $\mathbf{X}$ through the sampling model. The expected gain in Shannon information given by the design $\mathbf{X}$ is

$$E_{\mathbf{y},\boldsymbol{\theta}}\{\log[\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})/\pi(\boldsymbol{\theta})]\} = \int\int \log \frac{\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})}{\pi(\boldsymbol{\theta})} f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}d\mathbf{y}. \quad (1)$$

This utility function is appropriate when the purpose of the experiment is inference about $\boldsymbol{\theta}$. Note also that (1) is the expected Kullback–Leibler distance between the prior and posterior distributions. Therefore, because the prior does not depend on the design, the design $\mathbf{X}$ maximizing (1) is the one that maximizes the utility function

$$U(\mathbf{X}) = \int\int \log[\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})] f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}d\mathbf{y}. \quad (2)$$

In the usual normal linear regression model with known error variance σ^2, maximizing (2) reduces to finding the design matrix $\mathbf{X}$ that maximizes $\det(\mathbf{X}^T\mathbf{X} + \mathbf{R}^{-1})$, where $\sigma^2\mathbf{R}$ is the conditional prior covariance matrix of normally distributed $\boldsymbol{\theta}$ given σ^2 and $\mathbf{R}$ is known. Although the procedure we will describe can easily be applied using numerous other Bayesian design criteria (see Chaloner and Verdinelli 1995), we restrict our consideration here to finding those designs that nearly maximize (2). This article refers to those designs that nearly maximize (2) as near-optimal Bayesian expected information gain (EIG) designs.

As mentioned earlier, the main problem with (2) is that, for many practical problems, the integration is intractable, and numerical methods are needed to find optimal Bayesian designs. This is also the case here. We propose a two-stage iterative process for finding near-optimal Bayesian EIG designs. The process is illustrated in Figure 1. In Stage 1 we use a GA to generate potentially high EIG designs, and in Stage 2 we use Monte Carlo simulation to numerically estimate the EIG utility of each of the candidate designs proposed in Stage 1.

2.1 Stage 1: Genetic Algorithm

A GA operates on a "population" of candidate "solutions" to the optimization problem. Traditional GAs consider a solution to be a bitstring (i.e., binary string), or chromosome, and the population is comprised of chromosomes having the same length and structure. In the case of experimental designs, a single chromosome completely defines an experimental design $\mathbf{X}$. Each chromosome is first divided into as many bitfields as there are observations or runs, and each bitfield is then further subdivided into sub-bitfields, which code the factor values for that run. Thus, each chromosome will have a length equal to the sum

of the number of bits needed to code each factor value multiplied by the number of runs in the experiment. For example, an n-run design involving three factors each taking on 2^k values in their respective ranges would require chromosomes of length $3n2^k$ bits.

In many GA applications, however, natural (base 10) coding of variables can be used without the necessity of resorting to binary coding. Because of the convenience of using natural variables, we likewise use this representation here. Thus, the length of each chromosome is simply the product of the number of runs and the number of factors. For convenience, suppose that the experiment of interest contains p factors, and n runs must be made. Thus, both p and n are fixed and known. If we consider each factor level as a gene, then each chromosome (or design $\mathbf{X}$) has np genes whose values we seek.

We describe the construction of an initial population of solutions and subsequent populations of solutions obtained by use of the genetic operators of crossover and mutation within the context of an elitist GA described in the following.

Recall that we have restricted our consideration to continuous factors. We further assume that the design region is bounded, say, $L_i \leq x_i \leq U_i$ for the ith factor x_i. To begin the GA process, we first generate an initial population of M random designs using independent uniform random draws for each of the p factors x_i for each of the n runs. We then evaluate the utility (the "fitness" in GA terminology) of each of these M random designs using the Stage 2 approach described in Section 2.2. These M designs are then ranked according to their utility—that is, designs with higher utility get lower ranks. This completes the first generation of the GA.

The second (and subsequent) GA generations are now populated using genetic crossover and mutation. First, consider genetic crossover. Two parent designs are randomly selected without replacement from the initial population with probability inversely proportional to the rank of their utility among all the M designs, making designs with higher utility more likely to be selected. Then, the n runs of the parent designs are paired, and a new crossover design is obtained by randomly selecting a run from each pair. The two parents are then returned to the initial population before the next crossover operation is performed. In this way, additional M designs are constructed using the crossover operator and, again, the utility of each new design is evaluated as in Stage 2.

For each of the initial M designs in the current population, we next apply genetic mutation to each factor x_i of each of the n runs. We also decrease the probability that mutation occurs as the number of generations increases. We accomplish the evolutionary phenomenon known as "punctuated equilibrium" (or periodic upsets) by executing the GA in successive batches of G generations which will be described in more detail later. For example, we often set $G = 100$.

It is desired to mutate each factor value with probability that decays exponentially as a function of generation. That is, mutations become less and less likely as the number of generations increases. To accomplish this, at generation g each factor value is mutated with probability $\exp(-\mu \times g)$ where μ is a user-specified mutation rate parameter.

Given that mutation of a factor value occurs, we then mutate the value with expectation approximately equal to the current value of the factor and variance that decreases with g. We accomplish this by means of a logit transformation as follows: first compute $z = (x - L)/(U - L)$ where x, L, and U are the current, minimum and maximum values of the factor. Then calculate $d = \log[z/(1 - z)] + [\text{uniform}(0, 1) - .5] \times \psi \times \exp(-\mu \times g)$. Here ψ is a user-specified parameter that controls the rate at which the variance decreases as a function of g. Finally, compute $u = L + (U - L) \times \exp(d)/[1 + \exp(d)]$ which is the desired mutated value between L and U. This logit transformation has the properties that the expected value is approximately equal to the current factor value x and the standard deviation decreases with g. Applying this mutation procedure to each of the initial M designs, we generate an additional M designs and the utility of each of these designs is calculated using the Stage 2 procedure.

In the original GA, each new population completely replaces the previous one. It can then happen that the best (most fit) solution in population $k + 1$ is worse than the best solution in population k. Consequently, very good solutions can be lost forever. A solution to this problem is to use an "elitist" GA. At each generation we keep the best M designs (those with highest utilities) out of the $3 \times M$ designs (M initial designs, M crossover designs and M mutated designs) which becomes the population of initial designs for the next generation.

We execute the above GA in batches of G generations in order to allow for "punctuated equilibrium." In simple terms, punctuated equilibrium is an observed genetic phenomenon in which mutations essentially decrease over time but with periodic upsets in this process (i.e., periodic large-scale catastrophic mutations are occasionally permitted to occur). The best M solutions after a given batch has been completed become the initial set of designs for the next batch of G generations (with g reset to 1 for each batch). Note that the probability of mutation is also reset to its original level with each new batch (and subsequently decreases with each new generation in a batch). After several batches of G generations of solutions have been obtained in this way, we finally report the design having the highest utility as our desired near-optimal Bayesian experimental design. An algorithmic description of this GA process is given in the Appendix. Section 3 illustrates the performance of this GA.

2.2 Stage 2: Utility Estimation

The utility for each of the GA-produced candidate designs generated at Stage 1 is estimated in Stage 2. For a given candidate design $\mathbf{X}$, we propose estimating the utility in (2) by Monte Carlo simulation. We assume that it is possible to sample the known prior distribution $\pi(\boldsymbol{\theta})$ and assumed sampling model $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$ conditional on $\boldsymbol{\theta}$ and $\mathbf{X}$. We consider two cases:

1. the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$ is available in closed form; and

2. the posterior distribution is unavailable in closed form.

If the posterior distribution is available in closed form, then we estimate the utility in (2) directly (using Monte Carlo simulation)

as

$$\hat{U}(\mathbf{X}) = \frac{1}{L}\sum_{l=1}^{L}\log[\pi(\boldsymbol{\theta}^{(l)}|\mathbf{y}^{(l)},\mathbf{X})], \tag{3}$$

where $\{(\boldsymbol{\theta}^{(l)},\mathbf{y}^{(l)}), l = 1,2,\ldots,L\}$ denote the L corresponding dependent pairs of randomly sampled values: $\boldsymbol{\theta}^{(l)}$ from the prior distribution $\pi(\boldsymbol{\theta})$, and $\mathbf{y}^{(l)}$ conditionally from the sampling model $f(\mathbf{y}|\boldsymbol{\theta}^{(l)},\mathbf{X})$.

If the posterior distribution is unavailable in closed form, then we cannot use (3) directly. In this case, we estimate (2) as

$$\hat{U}(\mathbf{X}) = \frac{1}{L}\sum_{l=1}^{L}\log\left[\frac{f(\mathbf{y}^{(l)}|\boldsymbol{\theta}^{(l)},\mathbf{X})\pi(\boldsymbol{\theta}^{(l)})}{\hat{f}(\mathbf{y}^{(l)}|\mathbf{X})}\right], \tag{4}$$

where $\{(\boldsymbol{\theta}^{(l)},\mathbf{y}^{(l)}), l = 1,2,\ldots,L\}$ is the same set of randomly sampled values as in (3). Here $\hat{f}(\mathbf{y}^{(l)}|\mathbf{X})$ is a suitable estimate of the marginal distribution evaluated at $\mathbf{y}^{(l)}$; that is, an appropriate estimate of the posterior normalizing constant for the given design $\mathbf{X}$.

Numerous methods have been proposed for estimating a normalizing constant. The more popular methods include the Laplace approximation and its variants (Tierney and Kadane 1986), Monte Carlo simulation methods such as importance sampling (Geweke 1989; Hammersley and Handscomb 1964), reciprocal importance sampling (Gelfand and Dey 1999), bridge sampling (Meng and Wong 1996) and path sampling (Gelman and Meng 1998). Two excellent surveys of existing methods are Gelman and Meng (1998) and DiCiccio, Kass, Raftery, and Wasserman (1997). In two examples we will consider in Section 3, the posterior is unknown, and we calculate $\hat{f}(\mathbf{y}^{(l)}|\mathbf{X})$ by numerically integrating the product of the sampling model and the prior distribution over $\boldsymbol{\theta}$. However, if the dimensionality of $\boldsymbol{\theta}$ exceeds three, numerical integration is generally infeasible and one of the other methods mentioned above must be used.

3. EXAMPLES

We now illustrate the performance of the two-stage iterative procedure in Figure 1 using five examples: single-factor quadratic regression, single-factor stylized quadratic regression, three-factor quadratic response surface, single-factor nonlinear regression, and single-factor logistic regression.

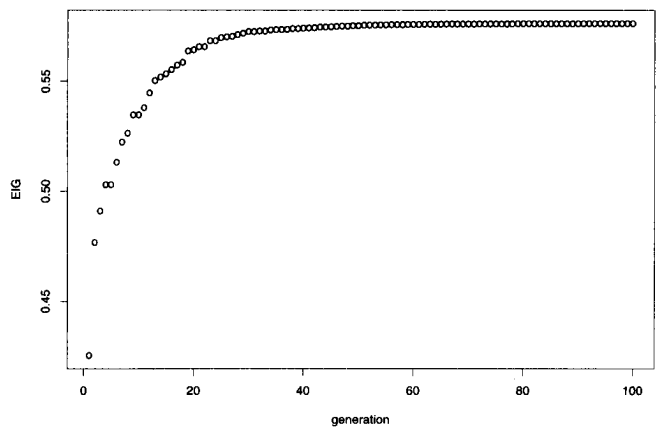

Figure 2. Example 1 EIG Evolutionary Plot

3.1 Example 1: Single-Factor Quadratic Regression

Pilz (1991, example 12.5) considered the quadratic regression model

$$\mathbf{y}|\boldsymbol{\theta},\mathbf{X} \sim N(\theta_0 + \theta_1\mathbf{X} + \theta_2\mathbf{X}^2, \sigma^2\mathbf{I}), \tag{5}$$

where $\mathbf{X}^T = (x_1, x_2, \ldots, x_n)$ denotes the n-vector of design values and $x_i \in [-1,1]$ and σ^2 is assumed known. The conditional prior distribution of $\boldsymbol{\theta}^T = (\theta_0,\theta_1,\theta_2)$ given σ^2 is $\boldsymbol{\theta}|\sigma^2 \sim N(\boldsymbol{\nu},\sigma^2\mathbf{R})$. Let $\boldsymbol{\nu}^T = (0,0,0)$, $R = \text{diag}(0.1,0.1,0.2)$ and $\sigma^2 = 1$. Because σ^2 is known, the EIG in (2) has a closed form and can be evaluated exactly. According to Pilz (1991), the optimal Bayesian D-optimal design (which as mentioned previously is the same as the optimal EIG design because σ^2 is known) places one half the design runs at -1 and one half at 1. This is a surprising result since the optimal design to estimate a second order model (three parameters) contains only two design points. Pilz (1991) provided a nice discussion of the related issue of one-point designs:

> The possibility of the existence of optimal one-point designs arises from the fact that the Bayesian information matrix is positive definite whatever the design. In paricular, there is good hope for the optimality of such designs if the prior precision matrix has a convenient structure, for example such that the prior knowledge arises from previous observations with a suitable (almost optimal) "prior" design.

For $n = 4$, then the optimal EIG design is $\mathbf{X}^T = (-1,-1,1,1)$.

We ran the GA described in Section 2 for this problem with the following parameters: $n = 4, p = 1, \mu = 0.01, \psi = 1, M = 10$, one batch with $G = 100$; namely, populations of size 10 are generated for 100 generations. The result obtained after this implementation of the GA is $\mathbf{X}^T = (-0.999892, -0.999993, 0.999905, 0.999854)$ with an EIG of 0.576272, where EIG is defined in (1). See Figure 2 for a plot which shows how the EIG increases over the 100 generations.

Pilz (1991) discussed the five-run case ($n = 5$) and stated that approximate exact designs could be obtained by rounding one half at -1 and 1 giving $\mathbf{X}^T = (-1,-1,-1,1,1)$ or $(-1,-1,1,1,1)$. Using two different starting seeds for the random number generators used in the GA, the following results were obtained: $\mathbf{X}^T = (-0.999923, -0.999912, 0.999909, 0.999890, 0.999864)$ with an EIG of 0.703678 and $\mathbf{X}^T = (-0.998929, -0.9989952, -0.998343, 0.999070, 0.999116)$ with an EIG of 0.702559. These results suggest that $(-1,-1,-1,1,1)$ and $(-1,-1,1,1,1)$ are indeed exact designs; that is, optimal designs for the five run case.

3.2 Example 2: Single-Factor Stylized Quadratic Regression

Dette (1993b) considered the following stylized quadratic regression model

$$\mathbf{y}|\boldsymbol{\theta},\mathbf{X} \sim N(\theta_1(1-\mathbf{X}) + \theta_2\mathbf{X}^2, \sigma^2\mathbf{I}), \tag{6}$$

where $\mathbf{X}^T = (x_1, x_2, \ldots, x_n)$ denotes the n-vector of design values, $x_i \in [0,1]$ and σ^2 is assumed known.

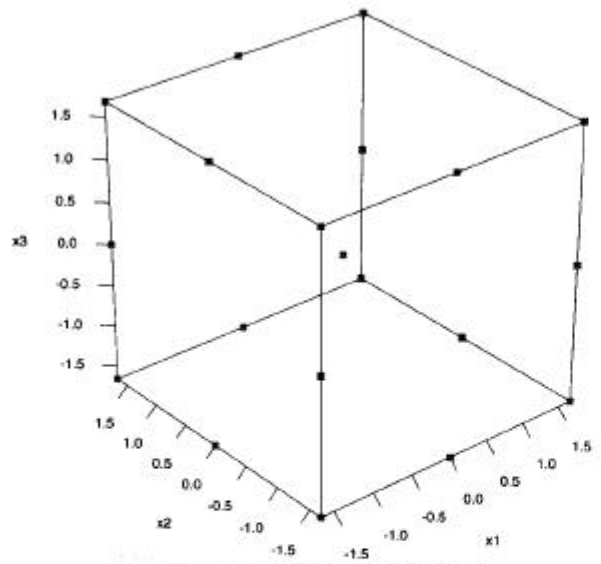

Figure 3. Example 3 Near-Optimal Design

The conditional prior distribution of $\boldsymbol{\theta}^T = (\theta_1, \theta_2)$ given σ^2 is $\boldsymbol{\theta}|\sigma^2 \sim N(\boldsymbol{\theta}_0, \sigma^2\mathbf{R})$. Let $\boldsymbol{\theta}_0^T = (0,0)$, $R = \text{diag}(0.1, 0.2)$ and $\sigma^2 = 1$. Because σ^2 is known, the EIG in (1) has a closed form and can be evaluated exactly. According to Dette (1993b), for $n = 8$, the optimal Bayesian D-optimal design (which is also the optimal EIG design because σ^2 is known) places six runs each at 0 and two runs at 1.

We ran the GA for this problem with the following parameters: $n = 8, p = 1, \mu = 0.001, \psi = 3, M = 10$, one batch with $G = 100$. The result obtained was (to 6 decimal places) the optimal design found theoretically by Dette (1993b); that is, (0,0,0,0,0,0,1,1) with an EIG of 1.182362.

3.3 Example 3: Three-Factor Quadratic Response Surface

Draper and Smith (1981, p. 390) considered an experiment on a chemical process involving three factors identified as being important. To better understand the impact of the factors on the response of interest, yield, a 20-run central composite design was performed. In coded variables, the central composite design used has eight cube points, one at each corner point of a cube of length 2 centered at (0,0,0), six star points consisting of a pair for each of the three axes which are on the axes 5/3 away from (0,0,0), and six center points at (0,0,0). The central composite design allows a quadratic response surface (regression) model to be fit

$$\mathbf{y}|\boldsymbol{\theta}, \mathbf{X} \sim N\left(\theta_0 + \sum_{i=1}^{3}\theta_i X_i + \sum_{i<j}\theta_{ij}X_iX_j + \sum_{i=1}^{3}\theta_{ii}X_i^2, \sigma^2\mathbf{I}\right), \quad (7)$$

where the design $\mathbf{X}$ is $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ whose values are in $[-5/3, 5/3]$.

Here, we assume that σ^2 is also unknown and use the conjugate prior for $(\boldsymbol{\theta}, \sigma^2)$, where $\boldsymbol{\theta}$ is the vector of 10 regression coefficients in (7). The so-called normal-inverse gamma prior has the following form: $\sigma^2 \sim IG(\alpha, \beta)$ and the conditional distribution of $\boldsymbol{\theta}$ given σ^2 is $\boldsymbol{\theta}|\sigma^2 \sim N(\boldsymbol{\nu}, \sigma^2\mathbf{R})$. Rather diffuse priors that contain the point estimates of the parameters based on data from the central composite design were specified as follows: $(\alpha = 6, \beta = 100)$ giving a prior mean and variance for σ^2 of 20 and 100, respectively, $\boldsymbol{\nu}$ is the zero vector, and $\mathbf{R}$ is the diagonal matrix of 5's except for the [1,1] entry corresponding to the intercept being 500.

To evaluate EIG in (1), while the posterior joint density of $(\boldsymbol{\theta}, \sigma^2)$ has a closed form (i.e., normal-inverse gamma), the integral in (1) does not. Hence, (1) is estimated using (3) in which $L = 10{,}000$ Monte Carlo simulations were performed.

For the GA, we have $n = 20, p = 3, \mu = 0.01, \psi = 1, M = 10$. The results from several batches totaling 1100 generations suggested the design displayed in Figure 3. In particular, the design points obtained had factor levels near $\pm 5/3$ or 0 so that using these exact values gave a slightly better design whose EIG is 26.253689 based on $L = 1{,}000{,}000$. Note the symmetry of the near-optimal design displayed in Figure 3 which shows one center point run and the remaining runs on the cube faces whose length is 10/3. Note the missing point on the upper right edge would be there if we considered a 21-point design. As a matter of comparison, the central composite design has a smaller EIG of 19.994195 also based on $L = 1{,}000{,}000$. These designs differ because the criteria used to obtain them are different. The classical central-composite design considers constant prediction variance at all points equidistant from the center, while the Bayesian design considers the expected information gain from the experiment. However, in both cases, the same response surface model is assumed.

3.4 Example 4: One-Factor Nonlinear Regression

Sebastiani and Wynn (2000) considered experimental design for a first-order decay nonlinear regression model:

$$\mathbf{y}|\theta, \mathbf{X} \sim N(\exp(-\theta\mathbf{X}), \sigma^2\mathbf{I}), \quad (8)$$

where $\mathbf{X}^T = (x_1, x_2, \ldots, x_n)$ denotes the n-vector of design values and $x_i \in [0, 1]$ and σ^2 is assumed known.

Here we consider the $n = 3$ case with $\sigma^2 = 0.25$. To illustrate the use of an asymmetric prior distribution, we took the prior distribution of θ to be right-triangular(1,3) whose density is $\pi(\theta) = (\theta - 1)/2$. Now the posterior of θ in EIG given in (1) does not have a closed form but can be approximated by

$$\frac{f(\mathbf{y}|\theta, \mathbf{X})\pi(\theta)}{\hat{f}(\mathbf{y}|\mathbf{X})}, \quad (9)$$

where $\hat{f}(\mathbf{y}|\mathbf{X})$ was obtained by a one-dimensional numerical integration.

Sebastiani and Wynn (2000) showed that the optimal design for a three-point discrete uniform prior at (1,2,3) was a one-point design with all three runs at 0.5628. Likely, the optimal design is also a one-point design for the right-triangular prior considered here. A heuristic optimization was employed that evaluated a number of one point designs that led to the near-optimal design with all three runs at 0.35625; its EIG is 0.138378 (based on $L = 100{,}000$) for purposes of comparison.

We ran the GA with $n = 3, p = 1, \mu = 0.01, \psi = 3, M = 10$, one batch with $G = 100, L = 10{,}000$ and obtained the design (0.37218, 0.36468, 0.39967). Its EIG based on $L = 100{,}000$

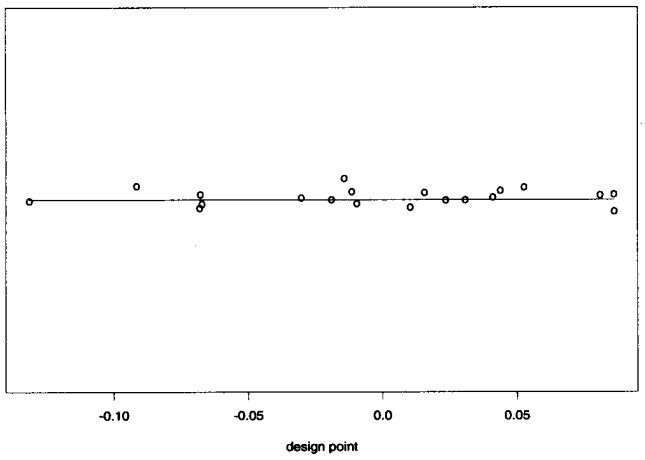

Figure 4. Example 5 Near-Optimal Design

is 0.137985. Thus, the GA found a design comparable to the near-optimal design given above.

3.5 Example 5: One-Factor Logistic Regression

Chaloner and Larntz (1989) considered experimental design for a logistic regression model:

$$\mathbf{y}|\boldsymbol{\beta}, \mu, \mathbf{X} \sim \text{Bernoulli}\left[1/(1+\exp(-\boldsymbol{\beta}(\mathbf{X}-\mu)))\right], \quad (10)$$

where $\mathbf{X}^T = (x_1, x_2, \ldots, x_n)$ denotes the n-vector of design values and $x_i \in [-1, 1]$.

Here we take the prior distribution for $\boldsymbol{\beta}$ to be Uniform$(6, 8)$, μ to be Uniform$(-0.3, 0.3)$ and $n = 20$. Now the posterior of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu)$ does not have a closed form but can also be obtained by (9) where the estimated marginal $\hat{f}(\mathbf{y}|\mathbf{X})$ was obtained by a two-dimensional numerical integration.

We ran the GA with $n = 20, p = 1, \mu = 0.01, \psi = 3, M = 10$, one batch with $G = 200, L =$ 10,000 and obtained the design displayed in Figure 4. (The design points have been jittered on the vertical scale to distinctly see all 20 design points.) Its EIG based on $L =$ 100,000 is 0.924180.

Chaloner and Larntz (1989) investigated another criterion, the expected log determinant of observed information, in which the optimal design was approximately $(-0.3, 0, 0.3)$ with weights (0.36,0.28,0.36) (read off their Figure 1). For $n = 20$, the number of runs at $(-0.3, 0, 0.3)$ are (7,6,7) and its EIG is 0.801324 based on $L =$ 100,000. As a matter of comparison, placing (7,6,7) runs at $(-1, 0, 1)$, approximately uniformly spread over the experimental region at three points, has an EIG of 0.520181. Even taking the near-optimal design displayed in Figure 4 and rounding to the nearest tenth (i.e., (5,11,4) runs at $(-0.1, 0, 0.1)$) has an EIG of 0.840404.

4. CONCLUSIONS

This article has shown how GAs can be used to find near-optimal Bayesian experimental designs. Here, we considered the expected Shannon information gain, but other design criteria can be handled easily. This methodology was illustrated with a wide range of examples. The methodology is easy to implement and allows a practical approach for designing even more complicated experiments. The near symmetry of the resulting best designs may suggest a symmetrical design which may indeed be optimal. In any case, the best designs found by GAs are likely to be good practical designs and can always be compared against designs suggested by the experimenter's intuition.

One possible modification to our approach would be to include in the starting set of designs to which the GA is applied the designs suggested by the experimenter's intuition. Although not illustrated in our examples, GA's can be applied to ANOVA models as well as regression models. In summary, we believe that GAs provide a useful addition to the statistical practitioner's toolkit for designing experiments.

APPENDIX

Pseudo-code for finding near-optimal Bayesian experimental designs via a Genetic Algorithm

Notation:

k - number of factors
n - number of runs
x_i - ith factor with range $L_i \le x_i \le U_i$
$\mathbf{X}$ - design, an $n \times k$ matrix
M - population size
G - number of generations

Pseudo-code:

Generation 0:

- generate M random designs $\mathbf{X}$ by drawing x_i from Uniform(L_i, U_i) for each of n runs and k factors
- evaluate utility (expected information gain) for each design (see (2) or the estimates in (3) or (4))
- order designs by decreasing utility

Perform generation g for generations $g = 1, \ldots, G$ {
Generation g

- generate M designs X by `CROSSOVER` (see below)
- generate M designs X by `MUTATION` (see below)
- evaluate utility of the $2M$ designs generated by crossover and mutation,
- order $3M$ designs ($2M$ designs and top M designs from generation $g-1$) by decreasing utility
- retain M designs with largest utility for generation $g+1$

}
`CROSSOVER`

- from the M designs retained from the previous generation, pick two designs with probability inversely proportional to their utility rank (largest utility has rank 1)
- the ith run of the generated design is generated by randomly choosing from the ith runs of the two picked designs, $i = 1, \ldots, n$

`MUTATION`

- for each of the M designs retained from the previous generation (referred to as current designs), a new design is generated as follows

- each entry of a current design is mutated with probability $\exp(-\mu \times g)$; that is, the mutation probability depends on g
- if an entry is mutated, the new entry is obtained by drawing from a particular distribution (see Section 2.1) which is approximately centered at the current entry and whose variance depends on the tuning parameter ψ.

[Received July 2000. Revised January 2001.]

REFERENCES

AndereRendon, J., Montgomery, D. C., and Rollier, D. A. (1997), "Design of Mixture Experiments Using Bayesian D-Optimality," *Journal of Quality Technology*, 29, 451–463.

Bernardo, J. M. (1979), "Expected Information As Expected Utility," *The Annals of Statistics*, 7, 686–690.

Broudiscou, A., Leardi, R., and Phan-Tan-Luu, R. (1996), "Genetic Algorithm as a Tool for Selection of D-Optimal Design," *Chemometrics and Intelligent Laboratory Systems*, 35, 105–116.

Chaloner, K., and Larntz, K. (1989), "Optimal Bayesian Design Applied to Logistic Regression Experiments," *Journal of Statistical Planning and Inference*, 21, 191–208.

Chaloner, K., and Verdinelli, I. (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 273–304.

Chatterjee, S., Laudato, M., and Lynch, L. A. (1996), "Genetic Algorithms and Their Statistical Applications: An Introduction," *Computational Statistics and Data Analysis*, 22, 633–665.

Clyde, M., Muller, P., and Parmigiani, G. (1996), "Inference and Design Strategies for a Hierarchical Logistic Regression Model," in *Bayesian Biostatistics*, eds. D. A. Berry and D. K. Stangl, New York: Marcel Dekker, pp. 297.

DeGroot, M. H. (1962), "Uncertainty, Information and Sequential Experiments," *Annals of Mathematical Statistics*, 33, 404–419.

——— (1986), "Concepts of Information Based on Utility," in *Recent Developments in the Foundations of Utility and Risk Theory*, eds. L. Daboni, A. Montesano, and M. Lines, Dordrecht: Reidel.

Dette, H. (1993a), "Elfving's Theorem for D-Optimality," *The Annals of Statistics*, 21, 753–766.

——— (1993b), "Bayesian D-Optimal and Model Robust Designs in Linear-Regression Models," *Statistics*, 25, 27–46.

——— (1996), "A Note on Bayesian c- and D-Optimal Designs in Nonlinear Regression Models," *The Annals of Statistics*, 24, 1225–1234.

Dette, H., and Neugebauer, H. M. (1997), "Bayesian D-Optimal Designs for Exponential Regression Models," *Journal of Statistical Planning and Inference*, 60, 331–349.

Dette, H., and Sperlich, S. (1994), "A Note on Bayesian D-Optimal Designs for a Generalization of the Exponential-Growth Model," *South African Statistical Journal*, 28, 103–117.

——— (1996), "Some Applications of Stieltjes Transforms in the Construction of Optimal Designs for Nonlinear Regression Models," *Computational Statistics and Data Analysis*, 21, 273–292.

Dette, H., and Wong, W. K. (1998), "Bayesian D-Optimal Designs on a Fixed Number of Design Points for Heteroscedastic Polynomial Models," *Biometrika*, 85, 869–882.

Draper, N.R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: Wiley.

DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Flournoy, N. (1993), "A Clinical Experiment in Bone Marrow Transplantation; Estimating a Percentage Point of a Quantal Response Curve," in *Case Studies in Bayesian Statistics*, eds. C. Gatsonis, J. Hodges, R. E. Kass, and N. Singpurwalla, New York: Springer, pp. 324–336.

Forouraghi, B. (2000), "A Genetic Algorithm for Multiobjective Robust Design," *Applied Intelligence*, 12, 151–161.

Gelfand, A., and Dey, D. (1999), "Bayesian Model Choice: Asymptotic and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501–514.

Gelman, A., and Meng, X. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1340.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, New York: Addison-Wesley.

Guttman, I. (1971), "A Remark on the Optimal Regression Designs with Previous Observations of Covey-Crump and Silvey," *Biometrika*, 58, 683–685.

Haines, L. M. (1998), "A Class of Equivalent Problems in Statistics and Operations Research," *South African Statistical Journal*, 32, 43–66.

Hammersley, J., and Handscomb, D. (1964), *Monte Carlo Methods*, London: Chapman and Hall.

He, Z., Studden, W. J., and Sun, D. (1996), "Optimal Design for Rational Models," *The Annals of Statistics*, 24, 2128–2147.

Holland, J. M. (1992a), *Adaptation in Natural and Artificial Systems*, Cambridge, MA: MIT Press.

——— (1992b), "Genetic Algorithms," *Scientific American*, July, 66–72.

Lindley, D. V. (1956), "On the Measure of Information Provided By An Experiment," *The Annals of Statistics*, 27, 986–1005.

——— (1972), *Bayesian Statistics—A Review*, Philadelphia: SIAM.

Meng, X., and Wong, W. (1996), "Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Explanation," *Statistica Sinica*, 6, 831–860.

Michalewicz, Z. (1992), *Genetic Algorithms + Data Structures = Evolution Programs*, New York: Springer-Verlag.

Mukhopadhyay, S., and Haines, L. (1995), "Bayesian D-Optimal Designs for the Exponential Growth Model," *Journal of Statistical Planning and Inference*, 44, 385–397.

Muller, P. (1998), "Simulation Based Optimal Design," *Bayesian Statistics*, 6, 1–13.

Pilz, J. (1991), *Bayesian Estimation and Experimental Design in Linear Regression*, New York: Wiley.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Boston: Division of Research, Harvard Business School.

Sebastiani, P., and Wynn, H. P. (2000), "Maximum Entropy Sampling and Optimal Bayesian Experimental Design," *Journal of the Royal Statistical Society*, Ser. B., 62, 145–157.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423; 623-656.

Sinha, B. K. (1970), "A Bayesian Approach to Optimum Allocation in Regression Problems," *Calcutta Statistical Association Bulletin*, 19, 45–52.

Song, D., and Wong, W. K. (1998), "Optimal Two-Point Designs for the Michaelis-Menten Model with Heteroscedastic Errors," *Communications in Statistics-Theory and Methods*, 27, 1503–1516.

Smith, A. F. M., and Verdinelli, I. (1980), "A Note on Bayesian Design for Inference Using a Hierarchical Linear Model," *Biometrika*, 67, 613–619.

Stone, M. (1959), "Application of a Measure of Information to the Design and Comparison of Regression Experiment," *Annals of Mathematical Statistics*, 30, 55–70.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Verdinelli, I. (2000), "A Note on Bayesian Design for the Normal Linear Model with Unknown Error Variance," *Biometrika*, 87, 222–227.